ESTIMATION OF THE DIFFERENCE OF THE MEANS OF ANTI-SARS-COV-2 IGG AND IGM ANTIBODY LEVELS OF RECOVERED PATIENTS AT THEIR DISCHARGE: A RATIO TYPE IMPUTATION PROCEDURE

Carlos N. Bouza* , Agustin Santiago**, Jose M. Sautto**
*Department of Mathematics and Computation, University of Havana, Havana, Cuba
**Unidad Académica de Matemática, Universidad Autónoma de Guerrero, Mexico.

ABSTRACT
Missing data appears in the study of recovered COVID19-patients. They should be imputed for estimating adequately difference of means. A new predictor is developed and its variance is obtained. Numerical studies are developed using data on anti-SARS-CoV-2 IgG and IgM antibody levels.

KEYWORDS: imputation, difference, approximated variance, COVID19, Jacknife, Bootstrap,

MSC: 62D05

RESUMEN
Data faltante aparece en el estudio de pacientes recuperados de COVID19. Estos deben ser imputados para estimar adecuadamente la diferencia de medias.Un nuevo predictor es desarrollado y su varianza es obtenida. Estudios Numéricos se llevaron a cabo usando n data de los niveles de los antecuerpos anti-SARS-CoV-2 IgG e IgM.

PALABRAS CLAVE: imputación, diferencia, aproximación de la variancia, COVID19, Jacknife, Bootstrap

1. INTRODUCTION

Researchers need testing the effectiveness of medical treatments, of agricultural pest controls developing experiments. They must face to deal with missing data in the sample due to failing in obtaining measurements of some of the experimental units. Incompleteness is identified as non-response. The researchers may be interested in the value of the variable of interest in the non-respondents for using appropriately statistical data analysis tools. The lack of full information inferring on the population parameters may be spoiled.

Estimating the difference of the means of two variables Y and X is common in many real-life problems. For example an epidemiologist is interested in the difference between the anti-SARS-CoV-2 IgG and IgM antibody levels, ecologists want to estimate the difference between the abundance of a pest after and before introducing a biological control of it, a social researcher observes some discussion forum in two occasions and is interested in the percentage of positive criteria, etc.

Commonly is needed to deal with the existence of missing data is problem. Rubin (1976) determined the existence of three of missingness mechanisms:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR).

In the case of MCAR and MAR the mechanisms are ignorable missingness mechanism. MCAR assumption may be difficult to be accepted but in laboratory studies experimented researchers may consider that in their study missingness is due to a chance mechanism.

An effective imputation method, under MCAR response mechanism, is derived as well as a resultant estimator. Some previous related papers are Al-Omari, A. I., Bouza -Herrera, C.N. (2013), Beaumont et al.. (2011), Berg et al. . (2016), Bouza-Herrera and Covarrubias-Melgar (2009).

Commercial softwares include some imputation methods in their library, see for example a: ICE, Imputation with Chained Equations (Stata), SAS; IVEware: Imputation and Variance Estimation Software, R Packages (MICE, Amelia, missForest, Hmisc, mi). See Raghunathan, et al. (2016), Waljee, et al. (2013), Rickert, (2016) .

The researcher should be able to fix some assumptions supporting that Y/X is similar to $\frac{Y_t}{X_t}$, for unit t. Then imputation may be considered as a good solution for predicting the missing observation of Y using the observed value of X. In the example on COVID19 anti-SARS-CoV-2 IgG and IgM antibody levels are standardized and a comparison a solution for dealing with predicting the difference between two measurements of them.

The behavior of the proposal is evaluated using resampling methods in a real-life study. Resampling procedures are replacing traditional statistical analysis approximation methods. They are computer–intensive and repeatedly resampling is used for inferring. Nowadays are available high speed and power in common computers which allow to perform inference using resampling.

Section 2 presents a proposal for imputing. It is a generalization of Liu (2006) and Bouza-Herrera and Covarrubias-Melgar (2009) proposals. The predictor of the difference is developed and its variance obtained. Section 3 presents a numerical study analyzing real life data where X=anti-SARS-CoV-2 IgG and Y=IgM antibody levels were measured in recovered patients. Missing data may be present in Y due to diverse causes. A data base of patients with full response used as an artificial population and missing data were artificially generated. The coverage provided by the application of Bootstrap and Jacknife confidence intervals was analyzed. The absolute relative mean error (approximation error) of the individual imputations was also analyzed.

2. **IMPUTATION OF THE NON-RESPONSES**

Datasets frequently must face the existence of missing values in statistical applications. When a data set contains missing values statisticians may decide using weighting or imputation methods. The decision depends on the nature of the non-response mechanism. Population surveys inevitably face the problem of dealing with incomplete data. Imputation may be used to complete datasets, by filling the missing values with adequate values. Imputation permits working with a complete data set, then the subsequent analyses may be implemented. Some papers on the uses of imputation in survey sampling are Al-Omari and Bouza - Herrera (2013), Yang-Kim (2018), Beaumont et al. (2011), Berg et al. (2016), Chen and Shen (2015), Liu et al. (2006)

Missing data arise in surveys as well as in clinical and epidemiological studies as well as in other scientific research. It is an important source of errors and in many cases invalidate using efficiently statistical tools for analyzing the data. This problem is the theme of a lot of papers and chapters in specialized books. Enlightening discussion may be obtained in Enders (2010), Good (2006), Haziza (2009).

The problem to be considered in this paper is the estimation of the difference between two variables X and Y:

$$\Delta \quad = \bar{X} - \bar{Y}$$

There are missing observations of Y. The random sample s is divided into

$$s(1) = \{units\ with\ full\ response\}, \|s(1)\| = n_1$$

$$s(2) = \{units\ with\ full\ response\ in\ X\ and\ missing\ measurments\ in\ Y\}, \|s(2)\| = n_2$$

Using the measurements obtained and imputing Y in the units in s(2). Consider to estimate the difference through

$$\widehat{\Delta} = \frac{n_1(\bar{x}_1 - \bar{y}_1) + n_2(\bar{x}_2 - \bar{y}_2{}^*)}{n}$$

The imputation is made, using the proposal of Liu (2006) and Bouza-Covarubias (2009) extension, as

$$y_i^* = \left(\frac{\sum_{t=1}^{n_1} r_{ty(1)}}{n_1}\right) x_i = \bar{r}_{y(1)} x_i;\ r_{ty(1)} = \frac{y_t}{x_t}$$

Then the imputed mean is

$$\bar{y}_2{}^* = \bar{r}_{y(1)}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} x_i\right)$$

For each unit in s(2) the expectation of the imputed value is

$$E\left(\frac{y_t}{x_t}x_i\right) = Cov\left(\frac{y_t}{x_t}, x_i\right) + E\left(\frac{y_t}{x_t}\right)E(x_i)$$

The use of independent random sampling sustains accepting that

$$E\left(\frac{y_t}{x_t}x_i\right) = E\left(\frac{y_t}{x_t}\right)E(x_i) = E\left(\frac{y_t}{x_t}\right)\bar{X}$$

Take

$$\varepsilon_{0t} = \left(\frac{y_t}{\bar{Y}} - 1\right);\ \varepsilon_{1t} = \left(\frac{x_t}{\bar{X}} - 1\right)$$

They have expectation zero and

$$E(\varepsilon_{0t})^2 = \frac{\sigma_y^2}{\bar{Y}^2};\ E(\varepsilon_{1t})^2 = \frac{\sigma_x^2}{\bar{X}^2};\ E(\varepsilon_{0t}\varepsilon_{1t}) \quad = \rho_{x,y}\frac{\sigma_y}{\bar{Y}}\frac{\sigma_x}{\bar{X}}$$

Assuming that

$$|\varepsilon_{1t}| < 1, \varepsilon_{1t}^q \to 0\ when\ g \uparrow$$

is possible to develop the Binomial expansion of $(1 + \varepsilon_{1t})^{-1}$. Then ,

$$y_t^* = y_t\left(\frac{\bar{X}}{x_t}\right) = \bar{Y}(1 + \varepsilon_{0t})(1 + \varepsilon_{1t})^{-1} = \bar{Y}(1 + \varepsilon_{0t} - \varepsilon_{1t} + \varepsilon_{1t}^2 - \varepsilon_{0t}\varepsilon_{1t} + O(\varepsilon_{1t}))$$

Its expectation is approximately

$$E(y_t^*) = \left[\left(\frac{\sum_{t=1}^{n_1}\bar{Y}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right) + O(\varepsilon_{1t})}{\bar{X}n_1}\right)\bar{X}\right] \cong \bar{Y}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)$$

As a result

$$E(\bar{y}_2{}^*) \cong \bar{Y}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)$$

Note that the prediction of the mean of Y is given by

$$\bar{y}_{imp} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}$$

Its conditional expectation is

$$E\left(\bar{y}_{imp}|n_2\right) = \frac{n_1\bar{Y} + n_2\bar{Y}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)}{n} = \bar{Y} + \frac{n_2\bar{Y}\left(C_x^2 - \rho_{x,y}C_xC_y\right)}{n}$$

The expectation of $E\left(\bar{y}_{imp}|n_2\right)$ is easily derived, considering that W_1 is the probability of a full response. It is given by

$$E\left(E\left(\bar{y}_{imp}|n_2\right)\right) \cong \bar{Y} + \frac{W_2\bar{Y}\left(C_x^2 - \rho_{x,y}C_xC_y\right)}{n}, W_2 = 1 - W_1$$

The proposed estimator of the difference is

$$\hat{\Delta}_{imp} = \frac{n_1(\bar{x}_1 - \bar{y}_1)}{n} + \frac{n_2(\bar{x}_2 - \bar{y}_2^*)}{n} = \bar{x} - \frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}$$

Then the following result holds:

Lemma: $Bias\left(\hat{\Delta}_{imp}\right) \cong -\frac{W_2\bar{Y}\left(C_x^2 - \rho_{x,y}C_xC_y\right)}{n}$ □

Clearly if the response probability is high the bias is negligible. That is
$Bias\left(\hat{\Delta}_{imp}\right) \to 0\ \ when\ W_1 \to 1\ or\ n \to \infty.$

The error of the predictor is

$$\varphi\left(\hat{\Delta}_{imp}\right) = E\left(V\left(\hat{\Delta}_{imp}|n_2\right)\right) + V\left(E\left(\hat{\Delta}_{imp}|n_2\right)\right) = \varphi_1 + \varphi_2$$

As

$$E\left(\bar{y}_{imp}|n_2\right) \cong \bar{Y} + \frac{n_2\bar{Y}\left(C_x^2 - \rho_{x,y}C_xC_y\right)}{n}$$

the variance of the conditional variance is

$V\left[E\left(\bar{y}_{imp}|n_2\right)\right] \cong \frac{W_1W_2(\bar{Y}\left(C_x^2 - \rho_{x,y}C_xC_y\right))^2}{n}$ (1)

Note that, if the number on non-responses is small, it is negligible. On the other hand

$$V\left(\hat{\Delta}_{imp}|n_2\right) = V(\bar{x}) + V\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}|n_2\right) - 2Cov\left(\bar{x}, \frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}\middle|n_2\right) = V(1) + V(2|n_2) - COV$$

The first term is the variance of X in s

$$V(1) = V(\bar{x}) = \frac{\sigma_x^2}{n} \tag{2}$$

The second term is

$$V(2|n_2) = V\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}|n_2\right) = \left(\frac{n_1}{n}\right)^2 \frac{\sigma_y^2}{n_1} + \left(\frac{n_2}{n}\right)^2 V(\bar{y}_2^*)$$

Because, due to the independence $Cov(\bar{y}_1, \bar{y}_2^*) = 0$. Note that

$$E((\bar{y}_2^* - \bar{Y})^2|n_2) \cong \frac{\bar{Y}^2}{n_2\bar{X}^2} E\left((\varepsilon_{0t} - \varepsilon_{1t} + \varepsilon_{1t}^2 - \varepsilon_{0t}\varepsilon_{1t})|n_2 \quad\right) \cong \frac{\bar{Y}^2}{n_2\bar{X}^2}\left(C_y^2 + C_x^2 - \rho_{xy}C_xC_y\right)$$

Then is possible to use the approximation

$$V(2|n_2) \cong \left(\frac{n_2}{n}\right)^2 V\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}|n_2\right) \cong n_1\frac{\sigma_y^2}{n^2} + n_2\frac{\bar{Y}^2}{n^2\bar{X}^2}\left(C_y^2 + C_x^2 - \rho_{xy}C_xC_y\right)$$

Its expectation is

$$E\left(V(2|n_2)\right) \cong W_1\frac{\sigma_y^2}{n} + W_2\frac{\bar{Y}^2}{n\quad \bar{X}^2}\left(C_y^2 + C_x^2 - \rho_{xy}C_xC_y\right)$$

The third term is

$$\begin{aligned} COV &= Cov\left(\bar{x}, \frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}\middle|n_2\right) \\ &= E\left(\frac{n_1^2}{n^2}(\bar{x}_1\bar{y}_1) + \frac{n_2\bar{x}_2}{n} \times \frac{n_1\bar{y}_1}{n} + \frac{n_1n_2\bar{x}_1\bar{y}_2^*}{n^2} + \frac{n_2^2\bar{x}_2\bar{y}_2^*}{n^2}\right) - E(\bar{x})E\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}\right) \\ &= C(1) + C(2) + C(3) + C(4) - C(5) \end{aligned}$$

where

$$C(1) = E\left(\frac{n_1^2}{n^2}(\bar{x}_1\bar{y}_1)\right) = \frac{n_1^2}{n^2}\left(\rho_{xy}C_xC_y + \bar{X}\bar{Y}\right)$$

$$E\left(C(1)\right) = \frac{W_1W_2\left(\rho_{xy}C_xC_y + \bar{X}\bar{Y}\right)}{n}$$

$$C(2) = E\left(\frac{n_2\bar{x}_2}{n} \times \frac{n_1\bar{y}_1}{n}\right) = \frac{n_2\bar{X}}{n} \times \frac{n_1\bar{Y}}{n}$$

$$E\left(C(2)\right) = -\frac{W_2W_1\bar{X}\bar{Y}}{n}$$

$$E\left(C(3)\right) = \frac{W_2}{n}\bar{X}\bar{Y} + \frac{W_1}{n}\quad \bar{X}^2\bar{Y}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)$$

$$C(4) \cong \frac{n_2}{n^2}\left[(\sigma_x^2 + \bar{X}^2)(\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)) + \bar{X}\left(\bar{Y}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)\right)\right]$$

$$E\left(C(4)\right) \cong \frac{W_2}{n}[(\sigma_x^2 + \bar{X}^2) + \bar{X}\bar{Y}]\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)$$

$$C(5) = E(\bar{x})E\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2^*}{n}\right) \cong \bar{X}\bar{Y}\left(\frac{n_1}{n} + \frac{n_2}{n}\left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)\right)$$

$$E\left(C(5)\right) \cong \bar{X}\bar{Y}\left(W_1 + W_2(1 + C_x^2 - \rho_{x,y}C_xC_y\right)$$

Therefore

$$E(COV) \cong \vartheta = \sum_{h=1}^{5}\vartheta_h$$

where

$$\vartheta_1 = \frac{W_2}{n}\left[\left[\left(\bar{X}\quad - \bar{Y}\right)^2 - \bar{Y}^2\right)\right]P\right]$$

$$\vartheta_2 = \frac{W_1W_2\left(\rho_{xy}C_xC_y\right)}{n}$$

$$\vartheta_3 = \frac{W_2}{n}\quad \bar{X}^2\bar{Y}P$$

$$\vartheta_4 = \frac{W_2}{n} P\sigma_x^2$$
$$\vartheta_5 = \frac{W_2}{n}\bar{X}\bar{Y} - 2\bar{X}\bar{Y}(W_1) = \bar{X}\bar{Y}\left(\frac{W_2}{n} - 2W_1\right)$$
$$P = \left(1 + C_x^2 - \rho_{x,y}C_xC_y\right)$$

Then is proved the next theoretical result.

$Lemma$: $E\left(V\left(\hat{\Delta}_{imp}|n_2\right)\right) \cong \frac{\sigma_x^2}{n} + W_1\frac{\sigma_y^2}{n} + W_2\frac{\bar{Y}^2}{n\ \bar{X}^2}\left(C_y^2 + C_x^2 - \rho_{xy}C_xC_y\right) + \vartheta$□

Note that if W_1 is large

$$E\left(V\left(\hat{\Delta}_{imp}|n_2\right)\right) \cong \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}$$

The estimation of the derived sampling error is rather complicated. The use of a resampling methods is the most common solution when dealing with such problems in sampling survey applications. See for example Kolenikov (2010), Haziza (2009), Enders (2010), Chen and Shen (2015), Beaumont et al. (2011). In the next section are developed resampling procedures for coping with the estimation of $\left(V\left(\hat{\Delta}_{imp}|n_2\right)\right)$.

3. RESAMPLING PROCEDURES FOR THE ESTIMATION OF $\left(\boldsymbol{V}\left(\hat{\Delta}_{\boldsymbol{imp}}|\boldsymbol{n_2}\right)\right)$

When imputation is present must be described the effect of imputing missing data. In this section is presented a resampling based study of real-life data. The study serves for illustrating the effectivity of estimating the variance of the proposed estimator in determining confidence intervals. The intervals are estimated using Jackknife and Bootstrap, both are resampling procedures with a robust behavior. A comparison of their robustness is developed under various scenarios. The efficiency of the individual imputations is also analyzed. The use of resampling in real life are discussed in Righi et al. (2014), Chen and Shen (2015), Shao (2003) for example.

The evaluation of the behavior of the proposed imputation method is analyzed by estimating the coefficient intervals using Jacknife and Bootstrap. The coverage probability of them is estimated.

The effect of the imputation in approximating the missing data is also evaluated. The efficiency, in terms of the approximations of the imputed values, is analyzed by comparing the mean of the relative absolute errors observed in the Monte Carlo experiments.

Records of recovered patients with COVID-19 were obtained. The epidemiology protocols established a regular follow-up and observation of them. The interest is analyzing retrospectively the clinical characteristics of the recovered patients for evaluating the effect of serum-specific antibody levels on positiveness. The difference of the means of X=anti-SARS-CoV-2 IgG and Y=IgM antibody levels of recovered patients at their discharge. The data consisted of 370 recovered patients positive to COVID19. Following the ideas of Quatember (2016) these entries conformed an artificial population. Hence, in the study the involved parameters were known. The Monte Carlo experiments were based on M=1000 iid samples. Using the population values, samples of size n were selected independently. Bernoulli experiments generate n_2 entries to be considered as missing-data.

3.1. The Jacknife numerical experiment

The Jackknife is a resampling method which was proposed by Quenouille in (1949), see Good (2006). Nowadays, it is used commonly for variance estimation. Jackknife uses a group of observations of size n-d, d=1,..,n-1, from the sample. Tukey in 1958, see Good (2006).. , introduced the term "pseudovalue" for the outputs obtained from each group of size n-d. They are iid random variables and are used to obtain a simple estimator of the variance. Its efficiency in variance estimation in sampling is discussed extensively in Rao and Shao (1992), Kovar and Chen (1994).

In the developed Monte Carlo experiment, M samples are selected independently and the jackknife procedure is used for determining confidence intervals. The coverage probability and the mean of the relative absolute approximation error, provided by the method, are estimated. A pseudo code is described as follows:

Jacknife study of a population for $\theta = \Delta$ using $\hat{\Delta}_{imp}$

```
Input M, d, n, r=0, b=0, $\vec{Z} = (z_1, \dots, z_N)$; $z_i = (X_i, Y_i)\,\tau = 0, \delta = 0, \theta$
    Step (i). Generate $n_2$ with distribution B(n,$W_2$ ).
```

```
            While $r \leq \binom{n}{d}$
                Select a sample and determine $\vec{z} = (z_1, \dots, z_n)$ from $\vec{Z} = (z_1, \dots, z_N)$.
                Generate n$_2$ independent Bernoulli variables with parameter W$_2$
                If $\beta_t = 1\ then\ \ z_t = z_t^*$
                Arrange the imputed sample
                $\vec{z}_{imp,r} = \left((x_1, y_1) \dots, (x_{n_1}, y_{n_1}), (x_{n_1+1,} y^*_{n_1+1}), \dots, (x_{n_1+n_2,} y^*_{n_1+n_2})\right)$
                Compute the estimator $\hat{\theta}_{s_{jbr}}\left(\vec{z}_{imp,r}\right) = \hat{\theta}_{s_{jbr}}$ .
                Calculate $\delta_{imp,b} = \frac{1}{n_2} \sum_{t=n_1+1}^{n_1+n_2} \left|\frac{y_t - y_t^*}{y_t}\right|$
            r=r+1
            Step (ii) Calculate $\hat{V}\left(\hat{\theta}_{nb}\right) = \frac{1}{\binom{n}{d}} \sum_{r=1}^{\binom{n}{d}} \left(\hat{\theta}_{s_{jr}} - \bar{\theta}_{Jb}\right)^2$ ; $\bar{\theta}_{Jb} = \frac{1}{\binom{n}{d}} \sum_{s_{jr}} \hat{\theta}_{s_{jr}}$
            Step (iii) Calculate $IC_b(\theta) = \left(\bar{\theta}_{Jb} \mp 2\sqrt{\hat{V}\left(\hat{\theta}_{nb}\right)}\right)$ ; $\tau_b = \begin{cases} 1\ if\ \theta \in IC_b(\theta) \\ 0\ if\ not \end{cases}$,
                $\tau = \tau + \tau_b, \delta = \delta + \delta_{imp,b}$
        b=b+1
            Step (iv) If $b < M$ go to step (i)
        Calculate $estimated\ coverage = \hat{\gamma} = \frac{\tau}{M}$
        Calculate $estimated\ approximated\ absolute\ error = \hat{\varepsilon} = \frac{\delta}{M}$
END
```

The results of the Monte Carlo experiments are given below.
Table 1 presents the estimated coverage probabilities for 3 sample sizes and 3 probabilities of non-responses.

Table 1. $\hat{\gamma}$ in a **Jacknife study of a population for** Δ **using** $\widehat{\Delta}_{imp}$

		W$_2$=0,1			W$_2$=0,3			W$_2$=0,5	
d	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$
1	0,822	0,885	0,903	0,725	0,697	0,684	0,528	0,457	0,423
3	0,840	0,893	0,915	0,823	0,828	0,831	0,753	0,753	0,750
5	0,887	0,896	0,920	0,852	0,848	0,855	0,778	0,778	0,775

The results of Table 1 suggest that for smaller values of W$_2$ and larger values of d the coverage probabilities increase. It seems reasonable to argue that the jackknife CI`s derived by using $\widehat{\Delta}_{imp}$ perform better when the number of imputations is small and the number of pseudo values is large.
The mean of the relative approximation error analysis is presented in Table 2. The results sustain considering that the errors are smaller for smaller values of W$_2$ and that their values are stable for variations in the sample sizes. The increase in W$_2$ seems to have a small effect in $\hat{\varepsilon}$. Then, it may be considered that the imputations on Y are not seriously affected by increases by the sample sizes but by the number of missing observations.

Table 2. $\hat{\varepsilon}$ in a **Jacknife study of a population for** Δ **using** $\widehat{\Delta}_{imp}$

		W$_2$=0,1			W$_2$=0,3			W$_2$=0,5	
d	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$
1	1,392	1,384	1,400	1,952	1,944	1,953	1,943	1,937	1,939
3	1,391	1,188	1,107	1,919	1,917	1,917	1,921	1,926	1,915
5	1,347	1,117	1,104	1,560	1,568	1,554	1,476	1,470	1,475

3.2. The Bootstrap numeric experiment

Bootstrap is a powerful resampling method for estimating the distribution of estimators and test statistics. Bootstrap may be considered as a method for simulating the behavior of a statistical procedure from the empirical distribution derived from the observed data. Efron´s seminal paper proposed this nonparametric approach , see Efron (1982). It performs better than Jackknife in some circumstances. Under milder conditions Bootstrap provides good approximations to the distribution of statistics, coverage probabilities of confidence intervals, and rejection probabilities of tests. It is at least as accurate as the first-order asymptotic distribution obtained by classic approximation. Sometimes the bootstrap is more accurate than the Delta method. See discussions and application of Bootstrap in survey sampling in Shao (2003), Liu etal. (2006), Chen and Shen (2015).

A pseudo code is the presented below.

Bootstrap study of a population for $\boldsymbol{\theta} = \Delta$ using $\widehat{\Delta}_{imp}$

```
Input M, n, r=0, b=0, $\vec{Z} = (z_1, \ldots, z_N)$; $z_i = (X_i, Y_i) \tau = 0, \delta = 0, \theta, R, W_2$
    Step (i). Generate $n_2$ with distribution B(n,$W_2$ ).
    Step (ii): Draw the sample $\{\{z_i\}_{i=1}^{n}\}_r$ of size n=$n_1$ +$n_2$ with replacement (SRSWR) from the observed
        values $z_1$, …, $z_n$
        Generate $n_2$ independent Bernoulli variables with parameter $W_2$
        If $\beta_t = 1\ then\ z_t = z_t^*$
        Arrange the imputed sample
      $\vec{z}_{imp,r} = \left((x_1, y_1) \ldots, (x_{n_1}, y_{n_1}), (x_{n_1+1,} y_{n_1+1}^*), \ldots, (x_{n_1+n_2,} y_{n_1+n_2}^*)\right)$
        Determine a subsample of size n:
        $\{Z(u_i)_b\}_{i=1}^{n};\ \vec{z}_{br}^* = \left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\ , \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^*\right)_{br}$
            Compute the estimator
                $\hat{\theta}_{s_j}(\vec{z}_{br}^*) = \hat{\theta}_{s_{jbr}}$ .
            Calculate
                $\delta_{imp,br} = \frac{1}{n_2}\sum_{t=n_1+1}^{n_1+n_2} |y_t - y_t^*|$
        r=r+1
    Step (iii) while r<R go to Step (ii)
    Step (iv) calculate
  $V_{Boot}(g(\theta)) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{s_{jbr}} - \hat{\theta}_{Boot}^*\right)^2, \hat{\theta}_{Boot}^* = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{s_{jbr}}, \delta = \delta + \delta_{imp,br}$
    Step (v) Calculate
  $IC_b(\mathrm{g}(\theta)) = \left(\hat{\theta}_{Boot}^* = \mp 2\sqrt{V_B(g(\theta))}\right); \tau_b = \begin{cases} 1\ if\ \theta \in IC_b(\mathrm{g}(\theta)) \\ 0\ if\ not \end{cases}, \tau = \tau + \tau_b$
            $\tau = \tau + \tau_b, \delta = \delta + \delta_{imp,b}$
            b=b+1
        while b<M go to step (i)
        Calculate $estimated\ coverage = \hat{\tau} = \frac{\tau}{M}$
    Calculate $estimated\ approximated\ absolute\ error = \hat{\varepsilon} = \frac{\delta}{M}$
END
```

The results of the experimentations are given in Tables 3 and 4.

Table 3. $\hat{\gamma}$ in a **Bootstrap study of a population for Δ using $\widehat{\Delta}_{imp}$**

		W₂=0,1			W₂=0,3			W₂=0,5	
R	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$
20	0,892	0,895	0,925	0,902	0,905	0,908	0,902	0,907	0,918
30	0,901	0,913	0,929	0,906	0,917	0,923	0,928	0,928	0,931
50	0,927	0,946	0,948	0,919	0,921	0,929	0,930	0,947	0,947
100	0,948	0,947	0,962	0,951	0,954	0,955	0,955	0,955	0,971

The results of Table 3 suggest that for larger values of W_2 and larger values of R the coverage probabilities are closer to the goal value of 0,95. It seems reasonable to argue that the convergence of the parametric Bootstrap CI`s, derived by using $\widehat{\Delta}_{imp}$, is faster in terms of the number of bootstrap samples observed.

The analysis of the mean of the relative approximation error is presented in Table 4. The results sustain considering that the errors are smaller for smaller values of W_2 and that their values are stable for variations in the sample sizes. The increase in W_2 seems to have a small effect in $\hat{\varepsilon}$. Then, it may be considered that the imputations on Y are not seriously affected by increases by the sample sizes but by the number of missing observations.

Note that in the study of $\hat{\varepsilon}$, Jackknife and Bootstrap produced similar results. For R>30 the approximation errors are not affected seriously by changes in the other parameters.

Table 4. $\hat{\varepsilon}$ in a **Bootstrap study of a population for $\boldsymbol{\theta} = \Delta$ using** $\widehat{\Delta}_{imp}$

		W_2=0,1			W_2=0,3			W_2=0,5	
R	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$	$n = 20$	$n = 30$	$n = 50$
20	1,391	1,387	1,403	1,943	1,941	1,956	1,933	1,939	1,939
30	1,391	1,185	1,100	1,916	1,910	1,909	1,919	1,924	1,935
50	1,188	1,189	1,180	1,185	1,184	1,184	1,187	1,186	1,186
100	1,118	1,110	1,087	1,120	1,115	1,115	1,118	1,116	1,116

4. CONCLUSIONS.

The proposed imputation procedure was friendly to the epidemiologists, as it fitted with pre- conceived ideas on an acceptable way of substituting the missing data.

Bootstrap performed better in terms of the coverage probabilities of the CI`s.

The computational costs were very similar for Jackknife and Bootstrap methods.

RECEIVED: JANUARY, 2022.
REVISED: MARCH, 2022.

REFERENCES

[1] AL-OMARI, A. I., BOUZA -HERRERA, C.N. (2013): Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient, **Quality and Quantity** 47, 353-365.

[2] BEAUMONT, J.-F., H AZIZA , D. and B OCCI , C. (2011). On variance estimation under auxiliary value imputation in sample surveys. **Statist. Sinica** 21, 515–537.

[3] BERG, E., KIM, J. K. AND SKINNER, C. (2016). Imputation under informative sampling, J. Surv. Statist. Methodol. 4, 436–462.

[4] BOUZA-HERRERA, C.N, and D. COVARRUBIAS-MELGAR (2009): Estimating the difference of means with imputation of the missing observations. **Rev. Inv. Operacional** 30, 156-172.

[5] CHEN H. and Q. R. SHEN (2015):Variance Estimation for Survey-Weighted Data Using Bootstrap Resampling Methods: 2013 Methods-of-Payment Survey Questionnaire. **Technical Report No. 104 / Rapport technique # 104, Bank of Canada.**

[6] EFRON, B. (1982): **The Jackknife, the Bootstrap and Other Resampling Plans**. SIAM, Philadelphia.

[7] ENDERS, C. K. (2010): **Applied missing data analysis**. The Guilford Press,New York.

[8] GOOD, P.I. (2006). **Resampling methods,** 3 rd Ed., Boston : Birkhauser.

[9] HAZIZA , D. (2009): **Imputation and inference in the presence of missing data**. In Sample Surveys: Design, Methods and Applications (C. R. Rao and D. Pfeffermann, eds.). Handbook of Statist. 29 215–246. Elsevier, Amsterdam. MR26546

[10] KOLENIKOV, S. (2010): Resampling Variance Estimation for Complex Survey Data. **The Stata Journal** 10, 165–199.

[11] KOVAR. J. K. and E.J. CHEN (1994): Jacknife variance estimations of imputed survey data. **Survey Met.**, 45-52.

[12] LIU, L., YUJUAN T., YINGFU L. and G. ZOU (2006): Imputation for missing data and variance estimation when auxiliary information is incomplete. **Model Assisted Statistics and Applications**. 1 , 83–94.

[13] QUATEMBER, A. (2016): **The Generation of Pseudo-Populations**. Heidelberg: Springer.

[14] RAO J.N.K and J. SHAO (1992): Jackknife variance estimation with survey data under hot deck imputation. **Biometrika**, 79 , 811–822.

[15] RIGHI P. ,S. FALORSI and A. FASULO (2014): Methods for variance estimation under random hot deck imputation in business survey. **Rivista Di Statistica Ufficiale**, 1-2, 45-64.

[16] SHAO, J. (2003): Impact of the Bootstrap on Sample Surveys. **Statistical Science**, 18, 191-198.

[17] YANG, S. and KIM, J. K. (2018): Predictive mean matching imputation in survey sampling. **Statistics Preprints.** 139. https://lib.dr.iastate.edu/stat_las_preprints/139